What is the process of conducting a multiple regression power analysis using Stata for data analysis?

Authored by

stats writer

June 29, 2024

The process of conducting a multiple regression power analysis using Stata for data analysis involves several steps. First, the user must specify the sample size, effect size, and significance level for the analysis. Next, the data must be imported into Stata and organized into the appropriate format for multiple regression analysis. Once the data is prepared, the user can run the power analysis command in Stata, which will generate a power table with the desired statistical power and sample size. The results of the power analysis can then be interpreted to determine if the sample size is sufficient to detect the desired effect size. If the power is too low, the user can adjust the sample size or effect size to achieve the desired level of power. The process may need to be repeated multiple times until an appropriate balance between power and sample size is achieved. Overall, conducting a multiple regression power analysis using Stata is a crucial step in ensuring the validity and reliability of the results obtained from the data analysis.

Multiple Regression Power Analysis | Stata Data Analysis Examples

Introduction

Power analysis is the name given to the process for determining the sample size for a research study. The technical definition of power is that it is the probability of detecting a "true" effect when it exists. Many students think that there is a simple formula for determining sample size for every research situation. However, the reality is that there are many research situations that are so complex that they almost defy rational power analysis. In most cases, power analysis involves a number of

simplifying assumptions, in order to make the problem tractable, and running the
analyses numerous times with different variations to cover all of the contingencies.

In this unit we will try to illustrate how to do a power analysis for multiple
regression model that has two control variables, one continuous research variable and
one categorical research variable (three levels).

Description of the Experiment

A school district is designing a multiple regression study looking at the effect of
gender, family income, mother's education and language spoken in the home on the English language proficiency scores of Latino high school students. The variables gender and
family income are control variables and not of primary research interest. Mother's education
is a continuous research variable that measures the number of years that the mother attended
school. The range of this variable is expected to be from 4 to 20. The variable

language spoken in the home is a categorical research variable with three levels: 1) Spanish only, 2) both Spanish and English, and 3) English only. Since there are three levels, it will take two dummy variables to code language spoken in the home.

The full regression model will look something like this,

0

1

2

3

4

5

(homelang2)

Thus, the primary research hypotheses are the test of b3 and the joint test of b4 and b5. These tests are equivalent the testing the change in R2

when momeduc (or homelang1 and homelang2) are added last to the regression equation.

The Power Analysis

**We will make use of the Stata command power to do the power
analysis. To begin with, we believe, from previous research, that the R2 for the
full-model (r2f) with five predictor variables (2 control, 1 continuous research, and 2 dummy variables
for the categorical variable) will be will be about 0.48.**

**Let's start with the continuous predictor (momeduc). We think that it will add about 0.03 to the
R2 when it is added last to the model. This means that the R2 for the model
without the variable (the reduced model, r2r) would be about 0.45. The total number of
variables (ntested) is 5 and the number being tested (ncontrol) is 1. We will run
power command three times with power equal to .7, .8 and .9.**

```
power rsquared .45 .48, power(0.7) ntested(5)
```

```
ncontrol(1)

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0500
power = 0.7000
delta = 0.0577
R2_R = 0.4500
R2_F = 0.4800
R2_diff = 0.0300
ncontrol = 1
ntested = 5

Estimated sample size:

N = 187

power rsquared .45 .48, power(0.8) ntested(5)
ncontrol(1)
```

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0500
power = 0.8000
delta = 0.0577
R2_R = 0.4500
R2_F = 0.4800
R2_diff = 0.0300
ncontrol = 1
ntested = 5

Estimated sample size:

N = 228

power rsquared .45 .48, power(0.9) ntested(5) ncontrol(1)

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0500
power = 0.9000
delta = 0.0577
R2_R = 0.4500
R2_F = 0.4800
R2_diff = 0.0300
ncontrol = 1
ntested = 5

Estimated sample size:

N = 292

This gives us a range of sample sizes ranging from 187 to 292 depending on power.

Let's see how this compares with the categorical predictor (homelang1 & homelang2)
which uses two dummy

variables in the model. We believe that the change in R2 attributed to the
two dummy variables will be about 0.025. This would give an r2r of 0.455. The
ntested stays at 5 while the ncontrol is now 2.

```
power rsquared .455 .48, power(0.7) ntested(5) ncontrol(2)

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0500
power = 0.7000
delta = 0.0481
R2_R = 0.4550
R2_F = 0.4800
R2_diff = 0.0250
ncontrol = 2
ntested = 5
```

Estimated sample size:

N = 224

power rsquared .455 .48, power(0.8) ntested(5) ncontrol(2)

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0500
power = 0.8000
delta = 0.0481
R2_R = 0.4550
R2_F = 0.4800
R2_diff = 0.0250
ncontrol = 2
ntested = 5

Estimated sample size:

N = 273

```
power rsquared .455 .48, power(0.9) ntested(5) ncontrol(2)

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0500
power = 0.9000
delta = 0.0481
R2_R = 0.4550
R2_F = 0.4800
R2_diff = 0.0250
ncontrol = 2
ntested = 5

Estimated sample size:

N = 349
```

This series of power analyses yielded sample sizes ranging from 224 to 349. These sample

sizes are larger than those for the continuous research variable.

If it is the case that both of these research variables are important, we might want to take into that we are testing two separate hypotheses (one for the continuous and one for the categorical) by adjusting the alpha level. The simplest but most draconian method would be to use a Bonferroni adjustment by dividing the nominal alpha level, 0.05, by the number of hypotheses, 2, yielding an alpha of 0.025. We will rerun the categorical variable power analysis using the new adjusted alpha level.

. power rsquared .455 .48, power(0.7) ntested(5) ncontrol(2) alpha(.025)

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0250
power = 0.7000
delta = 0.0481
R2_R = 0.4550
R2_F = 0.4800
R2_diff = 0.0250
ncontrol = 2
ntested = 5

Estimated sample size:

N = 267

```
power rsquared .455 .48, power(0.8) ntested(5) ncontrol(2) alpha(.025)
```

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0250

power = 0.8000

delta = 0.0481

R2_R = 0.4550

R2_F = 0.4800

R2_diff = 0.0250

ncontrol = 2

ntested = 5

Estimated sample size:

N = 320

```
power rsquared .455 .48, power(0.9) ntested(5) ncontrol(2) alpha(.025)
```

Performing iteration ...

Estimated sample size for multiple linear regression

F test for R2 testing subset of coefficients

H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0250

power = 0.9000

delta = 0.0481

R2_R = 0.4550

R2_F = 0.4800

R2_diff = 0.0250

ncontrol = 2

ntested = 5

Estimated sample size:

N = 401

The Bonferroni adjustment assumes that the tests of the two hypotheses are independent which is, in fact, not the case. The squared correlation between the two sets of predictors is about .2, which is equivalent to a correlation of approximately .45. Using an internet applet to compute a Bonferroni adjusted alpha taking into account the correlation gives us an adjusted alpha value of 0.034 to use in the power analysis.

power rsquared .455 .48, alpha(0.034) power(0.7) ntested(5) ncontrol(2)

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0340
power = 0.7000
delta = 0.0481
R2_R = 0.4550
R2_F = 0.4800
R2_diff = 0.0250
ncontrol = 2
ntested = 5

Estimated sample size:

N = 248

power rsquared .455 .48, alpha(0.034) power(0.8) ntested(5) ncontrol(2)

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients

H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0340
power = 0.8000
delta = 0.0481
R2_R = 0.4550
R2_F = 0.4800
R2_diff = 0.0250
ncontrol = 2
ntested = 5

Estimated sample size:

N = 300

power rsquared .455 .48, alpha(0.034) power(0.9) ntested(5) ncontrol(2)

Performing iteration ...

Estimated sample size for multiple linear regression
F test for R2 testing subset of coefficients
H0: R2_F = R2_R versus Ha: R2_F != R2_R

Study parameters:

alpha = 0.0340
power = 0.9000
delta = 0.0481
R2_R = 0.4550
R2_F = 0.4800
R2_diff = 0.0250
ncontrol = 2
ntested = 5

Estimated sample size:

N = 378

Based on the series of power analyses the school district has decided to collect data on a sample of about 300 students. This sample size should yield a power of around 0.8 in testing hypotheses concerning both the continuous research (momeduc) variable and the categorical research variable language spoken in the home (homelang1 and homelang2). The nominal alpha level is 0.05 but has been adjusted to .034 to take into account the number of

hypotheses tested and the correlation between the predictors.

See Also